A Multimodal Genre Recognition Prototype

Bernabeu, J.F., Pérez-Sancho, C., Ponce de León, P.J., Iñesta, J.M., Calvo-Zaragoza, J.
University of Alicante
{*jfbernabeu, cperez, pierre, inesta, jcalvo*}*@dlsi.ua.es*

Abstract

In this paper, a multimodal and interactive prototype to perform music genre classification is presented. The system is oriented to multi-part files in symbolic format but it can be adapted using a transcription system to transform audio content in music scores. This prototype uses different sources of information to give a possible answer to the user. It has been developed to allow a human expert to interact with the system to improve its results. In its current implementation, it offers a limited range of interaction and multimodality. Further development aimed at full interactivity and multimodal interactions is discussed.

1. Introduction

In this paper, a multimodal and interactive prototype to perform music genre classification is presented.

Classification of music into different categories is an important task for retrieval and organisation of music libraries. In our team, several engines to solve this task have been developed. However, music genre recognition is a difficult task due to her subjective nature. Genre classification involves many aspects. For example, genre labels are inherently subjective and influenced by a number of cultural, art, and market trends. So perfect results can not be expected [3]. Moreover, the success rate can be different depending on the particular classifier and the data used to train and test the system. Nevertheless, the combination of several sources can improve the success rate, as shown in [6].

Obtaining descriptive features from an object from different information sources permits to perform a deeper and more informative description of it. A number of papers can be found in the literature where pattern recognition is based on multimodal information. In [9] the authors explain how multimodality in human interaction and multimedia information processing can help to improve the performance in different pattern recognition tasks, like manuscript text processing or gesture recognition from image sequences. In [4] the authors consider a video sequence as a multimodal information source, obtaining features of different nature from speech, audio, text, shapes, or colors. This approach works under an *early* scheme where features are combined in a compact representation for a single decision. Other approaches use a late scheme where various classifiers are utilized for the different information sources and are then combined into a decision. For example, in [5] a multiple classifier system for OCR is presented, based on hidden Markov models that provide individual decisions. The combination of them is performed with a voting system.

In the present work, we present a multimodal genre recognition GUI to help the user to make a decision in the difficult task of classifying a multi-track file MIDI in a given music genre. The GUI provides the user several classifiers from different data sources. Some of these classifiers use the information which is in the melody part. Hence, the GUI provides a tool to find out the track in which the main melody is. Finally, the user can combine the several classifiers to get a proper classification.

The next section brings a system overview, including descriptions of its core classification engines and auxiliary modules. Next, its current interaction capabilities are discussed, and finally, some conclusions and further development lines are presented.

2. System design

The multimodal genre recognition GUI consists of two main modules: the melody track selection (MTS) module and the genre classification (GC) module. The basic operation mode is described below. An user chooses a multi-track MIDI file which he wants to classify. Then, MTS module does the needed operations to return the track having the highest probability of being the melody. MTS module is described in section 2.1 in more detail. Once we have a melody track selected,

the flow of the information arrives to the GC module. The GC module needs a track to be labeled as melody, since some of the genre classification engines assume that the features are extracted from a melody line. The GC module is described in section 2.2 in more detail. Finally, the system returns the genre which has the highest probability.

After presenting the basic operations of the system we explain in more detail the different modules pointing out the machine learning techniques which are used by the different engines to make the decisions in the classification.

2.1. Melody track selection (MTS) module

The function of the MTS module is to help the user to make the decision of melody track selection. For this, we need to assume that, if the melody exists, it is contained in a single voice or track, and it is not changing among several tracks. This assumption is also taken by others authors [2], as there is empirical evidence that it is the case for much of today's symbolically encoded western music. At this point, the system needs an engine that gives the probability of each track to be the main melody. A possible strategy is to use the metadata information found in MIDI files. However, metadata present some drawbacks as for example, unreliability, subjectivity, and they can be missed. Another drawback of this approach is that such a method would obviously tell us nothing about the content of melody tracks. Hence, it was not considered here. Instead, a version of our melody track selector [10] was used for this task as described below.

First, empty tracks and tracks playing on the percussion channel (channel MIDI 10) are filtered out in this approach. Each remaining track is described by a vector of numeric descriptors extracted from the track content. Some features describe the track as a whole while others characterise particular aspects of its content. These descriptors are the input to a classifier that assigns to each track its probability of being a melody. A random forest classifier, an ensemble of decision trees, was chosen as the classifier. The WEKA[1] toolkit was used to implement the system.

There is a possibility that the MIDI file does not have a melody track. To solve this problem an additional track named "NO MELODY" with a heuristic fixed probability $p_0 = 0.22$ is added. Then, each probability track is re-normalized. So this p_0 acts as a threshold, in such a way that for any track i only if its $p_i > p_0$ is considered for being a melody. If $p_i \leq p_0$ for all tracks, a "NO MELODY" answer for the file is given.

[1] http://www.cs.waikato.ac.nz/ml/weka/

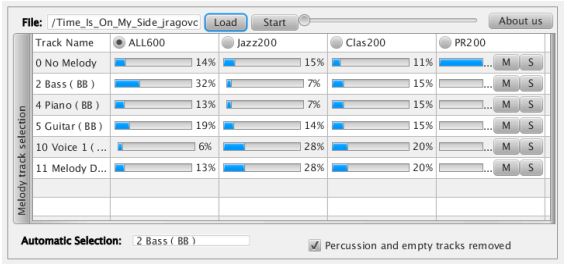

Figure 1. MTS module.

The GUI has several classifiers which were trained with different corpora. Specifically, four models were built using different data in the training phase. The files were downloaded from a number of freely accessible Internet sites. First, three corpora (JAZ200, CLA200, and PR200) made up of 200 files each, were created to set up the system and tune the parameter values. JAZ200 contains jazz files, CLA200 has classical pieces, and PR200 contains pop-rock songs. The other corpus named ALL600 is the union of these three corpora. The user can choose each model at any time selecting their radio buttons (see Fig 1). The right side shows the result, where each track gets its probability to be a melody displayed as a progress bar. Empty and percussion tracks are not showed by default, but the user have the option to see these tracks. Also, a slider control allows to listen to a specific section of the file and a mute/solo buttons are provided for each track.

2.2. Genre classification (GC) module

The function of the GC module is to help the user to make the decision of which genre corresponds to a target file. The working hypothesis is that melodies from a same musical genre may share some common low-level features, permitting a suitable pattern recognition system, based on statistical descriptors, to assign the proper musical genre to them. For this, it uses several engines that compute the probability to belong to a given genre. Now, the several genre classifiers are explained in more detail.

SVM based on melodic content features. The first classifier is a Support Vector Machine (SVM) classifier. The input data is based on statistical features of melodic content, like melodic, harmonic, and rhythmic descriptors.

There are 49 descriptors in total and they have been designed according to those used in musicological studies. For training the classifier each sample is represented as a labeled vector of statistical descriptors com-

puted from each melody segment available (see [8]). The SVM Weka implementation has been used to perform the SVM features classifier.

N-grams (notes). The second classifier is an N-gram classifier. The N-grams are used here as music words, that captures relevant information of the data and is suitable for a text categorization approach [7]. To do this we use a representation that combine pitch and note durations, using relative measures. The encoding method makes use of pitch intervals and inter-onset time ratios (IOR) to build series of symbols of a given length (N). There are two possible encodings, coupled (intervals and IOR are encoded together) and decoupled (separate symbols).

Once we have the MIDI information converted in a sequence of symbols, a language model is built from a training set of documents and used as classifier. For this, given a new, previously unseen, sequence of words, classification is done by selecting the class most likely to have generated that sequence. In this work, building and evaluation of the language models has been performed using the CMU SLM Toolkit[2], and a combination of both techniques, interpolation of models and the Witten-Bell discounting method have been used to solve the problem of the unseen samples. 4-grams models have been used here.

N-grams (chords) and metadata. Actually, this classifier can be seen as three classifiers: the first, N-grams (chords), using the chords provided by the harmonic structure of the music sequence; the second, Metadata, using the instrumentation information contained in a MIDI file metadata; and the third, "Combined", using an early combination of both data sources. In the three cases, the features give a single vector that will be the input to a classifier after a feature selection procedure.

Each file in the dataset is represented as a vector $x \in \{0,1\}^{H+I}$, where each component $x_i \in \{0,1\}$ codes the presence or absence of the i-th feature. H denotes the number of chords in the dictionary of possible harmonic combinations considered, H = 312 different chords in this work (see [7] for more details), and I is the number of possible instruments that, assuming the General MIDI standard for the sequence, will be 128 instruments plus 3 percussion sets. Therefore, I = 131.

There will be a probability of each feature associated to each class, depending on the frequencies found in the training set for the items in the classes. The decision will be taken combining these probabilities through a Naïve Bayes classifier. These classifiers are described in more detail in [6].

In order to select the features that contribute the most to class discrimination, a feature ranking has been established based on the Average Mutual Information (AMI) [1], that provides a measure of how much information about a class is able to provide a single feature.

Training set. Corpus 9GDB contains both melodic and harmonic information (including tonality). It consists in 856 files MIDI and Band-in-a-Box formats. It is divided in three musical genres: academic, jazz, and popular music. A second split of this database divides each genre in three subgenres, resulting in a total of 9 music subgenres.

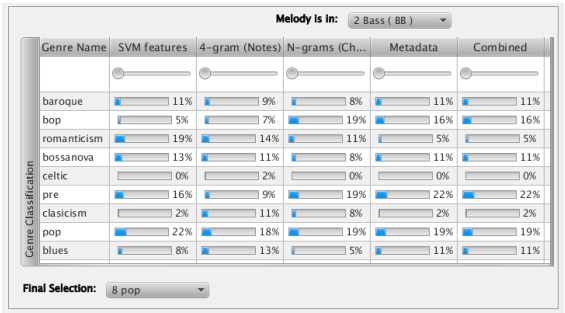

Figure 2. GC module.

This hierarchical structure allows the user to compare the classifiers at different levels, either at the first level with three broad genres, or at the second level with all nine subgenres, making the tool more versatile (see [7] for details).

Each classifier was trained with this corpus but each one provides the user different aspects to make a decision. As we explain above some of them uses the melody information and others the information contained in all the tracks or metadata. That is, each classifier uses as input different sources of information and can provide different answers for the same input file. In order to provide a mechanism to tune the final selection recommended by the system, the user can combine the classifiers assigning a weight for each model like a linear combination of the different classifiers.

3. User interaction

Music genre classification is clearly subjective and involves different aspects. Then, interaction with a human expert is needed to assess and validate the given answer by the different automatic systems. This interaction begins in the selection of which information

[2] http://www.speech.cs.cmu.edu/SLM/toolkit.html

the system uses and finishes in the validation or correction of the automatic classification. The goal is to minimize the number of interactions that a human expert should perform to obtain a reliable genre classification and when labeling a database of a number of MIDI files.

3.1. Interaction with MTS module

When the user works with the MTS module, he can hear the different tracks of the multi-part file and is provided to a mute/solo buttons to select the different tracks which he wants to hear when he is selecting the melody track. The user can see the probability of each track. Moreover, the user can select the several classifiers and can view or not the percussion and empty tracks.

3.2. Interaction with GC module

The main interaction with the GC module is to tune the final selection recommended by the system. The user can combine the classifiers assigning a weight for each model like a linear combination of the different classifiers. To do this each classifier have a slider bar to modify its weight in the final selection (see fig 2). Finally, the user has the option to change the selection recommended by the system if he considers that this selection is not proper.

4. Conclusions

In the current development state, this multimodal interactive music genre classifier prototype is capable of classifying multi-part music files. It can use several sources of information extracted from a MIDI file, such as melody features, melody notes, chords, and metadata information. The system allows the user to interact with both modules, MTS and GC, selecting and tuning the several classifiers involved.

This prototype is still in an early stage of development. It is conceived as a platform for interactive multimodal research in the context of symbolic music data. New features are planned for the near future, including: improved interface usability capabilities., addition of new source data input, such as audio multi-part files, addition of new user input modalitites, such as MIDI instrument live input, addition of new genre classifiers using different data sources, such as bass track or percussion track, addition of new classifiers based in different methods, such as tree grammars or tree automata.

The system can be extended to use the feedback user information. This way the classifiers could be trained incrementally with new samples classified by the user. Also, the system can provide a mechanism to save the classifier weights tuned by the user and to train them with user datasets allowing him to change the genre hierarchy.

Acknowledments. This work was supported by the projects DRIMS (TIN2009-14247-C02) and the PROMETEO/2012/017.

References

[1] T. M. Cover and J. A. Thomas. *Elements of information theory*. Wiley-Interscience, New York, NY, USA, 1991.

[2] A. Friberg and S. Ahlbäck. Recognition of the melody in a polyphonic symbolic score using perceptual knowledge. In *Proceedings of the 4th Conference on Interdisciplinary Musicology*, Thessaloniki, Greece, 2008.

[3] S. Lippens, J. Martens, M. Leman, B. Baets, H. Meyer, and G. Tzanetakis. A comparison of human and automatic musical genre classification. In *Proceedings of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, ICASSP 2004*, volume 4, pages 233–236, 2004.

[4] Z. Liu, Y. Wang, and T. Chen. Audio feature extraction and analysis for scene segmentation and classification. In *Journal of VLSI Signal Processing System*, pages 61–79, 1998.

[5] M. Liwicki and H. Bunke. Combining on-line and offline systems for handwriting recognition. In *Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on*, volume 1, pages 372–376, 2007.

[6] T. Pérez-García, C. Pérez-Sancho, and J. M. Iñesta. Harmonic and instrumental information fusion for musical genre classification. In *Proc. of. ACM Multimedia Workshop on Music and Machine Learning (MML 2010)*, pages 49–52, Florence (Italy), October 2010. ACM.

[7] C. Pérez-Sancho. *Stochastic Language Models for Music Information Retrieval*. PhD thesis, Alicante, Spain, July 2009.

[8] P. J. Ponce de León. *A statistical pattern recognition approach to symbolic music classification*. PhD thesis, Alicante, Spain, September 2011.

[9] G. Rigoll and S. Müller. Statistical pattern recognition techniques for multimodal human computer interaction and multimedia information processing. In *Information Processing, in Survey Paper, Int. Workshop Speech and Computer*, pages 60–69, 1999.

[10] D. Rizo, P. J. Ponce de León, C. Pérez-Sancho, A. Pertusa, and J. M. Iñesta. A pattern recognition approach for melody track selection in midi files. In T. A. Dannenberg R., Lemström K., editor, *Proc. of the 7th Int. Symp. on Music Information Retrieval ISMIR 2006*, pages 61–66, Victoria, Canada, 2006.